Computer aided diagnosis of Alzheimer's disease from brain images: a method robust to registration errors

Marta Sofia Aranha da Conceição
marta.conceicao@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2019

Abstract

Alzheimer's disease (AD) is a chronic progressive neurodegenerative disease affecting millions of people worldwide and prominently the elderly. While there is still no cure for AD, its early detection is crucial, as an effective management of the disease may help prevent the progression to more severe stages. The inherent uncertainty in the clinical diagnosis of AD has driven a search for biomarkers, where brain imaging, as positron emission tomography (PET), assumes a key role. For the acquired data to be used for computer aidded diagnosis of AD, however, it's usually necessary to perform image registration to a standard spatial coordinate system. This can be troublesome, as many challenges, including inter-subject anatomical variability, are encountered, so that possible misclassification errors might result from a poor coordinate transformation. In this work, in the attempt to find a method robust to such registration errors, and building from the texton-based approach, several methods were considered and applied on both registered and non-registered datasets. Other feature representations were considered, namely learned using a stacked sparse autoencoder (SSAE) and the raw voxel intensity values, either extracted from the whole brain, patches or identified regions of interest. Binary classification among cognitively normal subjects, AD and mild cognitive impairment (MCI) patients was performed in a dichotomous fashion and conclusions regarding the accuracy and robustness of the diferent methods were drawn, particularly confirming the robustness of the texton-based approach applied on the whole brain images and of the learned feature representations using SSAE.

Keywords: Alzheimer's disease; Computer aided diagnosis; PET; Textons; Registration; Classification

1. Introduction

Alzheimer's disease (AD) is an ultimately fatal neurodegenerative disease, and the most common cause of dementia, affecting millions of people worldwide and more prominently the elderly [1]. It's usually categorized in early onset familial AD, with well-defined genetic causes, or late onset sporadic AD, arising from genetic and environmental risk factors, some of which modifiable [2–4]. To date, two hallmarks of the disease have been established, namely plaques of accumulated beta-amyloid ($A\beta$) outside neurons, and neurofibrilliary tangles of hyperphosphorylated protein tau inside neurons [1, 2, 4–6]. Symptoms include difficulties with memory, language, problem-solving and other cognitive skills that affect the ability to perform daily activities, with final loss of bodily functions leading to death; nonetheless, these vary among AD patients and differences between early signs and age-related cognitive changes can be subtle [1, 3, 7].

Two early stages prior to symptom manifestation are recognized in AD, namely preclinical AD and mild cognitive impairment (MCI) [1, 2, 8]. In MCI, cognitive decline is greater than expected due to aging, but daily activities can still be performed without notable impairment, not fulfilling the criteria for dementia. It includes incipient AD, other causes of dementia and a form of cognitive impairment that doesn't progress to it, remaining stable or reverting to normal cognition [1,2,5]. As for the preclinical state preceding MCI, its existence is speculative, although evidence includes the presence of genetic risk factors, AD-like brain images, and abnormal cerebrospinal fluid (CSF) biomarkers in cognitively normal individuals [1].

Clinical diagnosis of AD can be based on medical and family history, neurological assessment, blood tests and neuroimaging, but none can be sufficiently reliable for the early stage detection desired for therapeutic intervention [1,2]. Concerning neuropsychological assessment, two tests that are usually performed are the Mini-Mental State Examination (MMSE), with a cut-off score for dementia of 26 (out of 30, lower values being observed in more severe stages), and the Clinical Dementia Rating (CDR), where 0 corresponds to no dementia, 0.5 to questionable or very mild, 1 to mild, 2 to moderate, and 3 to severe dementia, and MCI is often supported by a global rating of 0.5 on this scale [2, 8–11].

Given the uncertainty in clinical diagnosis of AD, neuroimaging, and namely FDG-PET, assumes a key role, constituting a robust biomarker of neurodegeneration where hypometabolism can be observed to precede cognitive symptoms and predict the rate of its decline in individuals who progress to AD with high sensitivity [1, 5, 12]. Indeed, despite providing a nonspecific indicator of metabolism that can be altered for other causes, as is-

chemia or inflammation, there is a characteristic ensemble of limbic and association regions typically hypometabolic in AD, found in neocortical association areas as the posterior cingulate, precuneus, temporoparietal and frontal multimodal association regions, as well as the hippocampus and medial temporal cortices, as illustrated in Figure 1 [12], while, in contrast to other dementia types, the primary visual cortex, sensorimotor cortex, basal ganglia and cerebellum are relatively unaffected [12–14]. The abnormalities found in AD with FDG-PET mirror those found with SPECT and MRI, and are reported to be more reliable for diagnostic purposes than these [13]. Less severe hypometabolism has been identified in MCI patients, some later converting to AD, such that it parallels cognitive function along the trajectory of normal, preclinical, prodromal, and established AD [12].

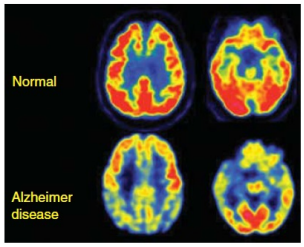

Figure 1: Example of transaxial FDG-PET images of cognitively normal vs. mild AD subjects.

Even if it outpaces therapeutic options, early detection of MCI is important as it can still have clinically significant implications, like slowing the rate of cognitive decline, reducing morbidity-affected life years, and improving quality of life [9]. As mentioned, neuroimaging tools can largely aid this diagnosis, and by allowing to automatize and prevent it from being affected by inter- and intra-rater reliability from the clinicians, computer aided diagnosis (CAD) of AD arises as a promising tool. Nonetheless, several processing steps are usually required to enable the acquired brain images to be used, one of them being image registration to a standard spatial coordinate system, so that each voxel corresponds to the same anatomical structure across all subjects [15]. This is a time consuming and troublesome step, particularly for deformable image registration, facing various challenges such as inter-subject anatomical variability. Many image registration algorithms are also available, adopting various deformation models and regularization strategies that give very distinct outcomes, possibly leading to misclassification errors, namely when there is prior knowledge of the regions of interest for diagnosis and this is affected by a poor coordinate transformation [15, 16]. In this work, in the attempt to develop a method for CAD of AD that doesn't require image registration, different approaches were applied to both registered and non-registered brain images, drawing conclusions regarding its efficiency, accuracy and robustness.

2. State of the Art

Several neuroimaging modalities have been used for CAD of AD, namely FDG-PET [16–18], sMRI [19–33], SPECT [34], DTI [35] and ASL [36], alone or combined for multimodal approaches [37–41].

Regardig feature extraction, studies have used either the whole brain [19,24,25,29,36,37,40,42], voxel-based approaches [16,23,27,28,32,36,38], regions of interest (ROIs) [16–18,20,24,26,28,30,31,34,35,39,41–43] or patches of the brain [21,32,33]. While the former two are intuitive, they either result in high dimensionality or disregard regional information. Using ROIs, in turn, representative features can be extracted from regions affected by AD, although these must be manually labeled or identified through feature selection. Alternatively, since the affected areas can be part of or span over multiple ROIs, patches of the brain can be used (combined hierarchically [21] or through feature concatenation [32]), which can efficiently handle the concerns of high feature dimensionality and sensitivity to small changes, by considering local information and allowing to extract richer information that might help improve diagnostic accuracy.

Different feature selection techniques have been applied, based on relevant statistics of the input data (as t-tests [21,22,24,27,28,32,38], voxel-based morphometry [28] and mutual information [16]), wrapper [16, 43] or else embedded methods [17, 18, 27, 32, 42]. Dimensionality reduction has also been attained using principal component analysis (PCA) [23, 24, 31, 33], partial least squares (PLS) [23], independent component analysis on means [29] or by imposing thresholds on voxel intensity [23].

As for the features commonly used for CAD of AD, these include voxel intensities and related statistics [16–18, 23, 28, 34, 36, 39, 41, 42], volume, shape, density and thickness of particular regions of the brain [21,22,30,31,36, 39, 42, 43]. Texture descriptors of the neuroimaging data have also been considered, including local binary patterns [44] and texton-based approaches [16], where the latter provides a full statistical representation of the responses to a predefined set of filters, turning the extracted image models into powerful descriptors.

Several deep learning strategies have emerged in recent years, providing high-level feature representations which can be more robust than hand-crafted ones. Amongst the approaches explored for CAD of AD that have achieved state of the art performances are convolutional neural networks (CNNs), usually combined with sparse autoencoders (SAE) to find the appropriate set of filters for the convolutional layer [19, 40], multi-instance convolutional neural networks [32], deep Boltzmann machines (DBM) [38], which, for multi-modal approaches, can fuse the complementary information without possible loss of correlation, and, similar to the autoencoder (AE), allow for reconstructing the input data from the hidden representations, while also resolving uncertainty about intermediate-level features and creating better data-dependent representations and statistics for learning, possibly outperforming it, and finally DenseNets, reported to achieve better performances than CNNs [33].

Since most studies perform voxel-wise comparisons or investigate brain abnormalities in ROIs, image registration is a common practice, namely to the Talairach or MNI space [45]. The former is based on a stereotaxic brain atlas constructed by the identification of given anatomical landmarks and brain segmentation into 12 subvolumes, where

a piecewise linear scaling is used to convert each individual brain image to this space [45, 46]. It has faced some criticism, not only because this scaling performs poorly comparing to nonlinear transformations, but mostly due to it being constructed from a single subject, hence not being representative of the neuroanatomy of the general population [45]. In light of this, an average brain template (MNI305) was put up, and later inspired the current standard MNI template (ICBM152), built by averaging 152 normal sMRI scans linearly matched to MNI305 [46, 47]. Despite this advantage provided by the MNI coordinate space, the Talairach brain atlas remains widely used in the literature, due to its construction from a set of anatomical landmarks [16, 20, 24, 31, 43].

Few studies have disregarded image registration, achieving this either only in the testing stage, or for both training and testing. Concerning the former, as performed in [27, 32], landmark-based feature extraction methods were applied, the landmarks being identified, in the training stage, on the template image, and, using their respective deformation fields, directly projected to the linearly-aligned training image. A shape-constrained random forest was then learned to non-linearly map different patches and its 3D displacement to the detected target landmarks, which could be applied in the testing stage to estimate a 3D displacement from every voxel to the potential landmark position, followed by majority voting. In [33], on the other hand, non-linear image registration was fully disregarded, for which patches within ROIs were identified from non-registered images and clustered using K-means, and a deep learning strategy was used to learn features that were further aggregated for region-level representations, followed by final image classification.

Regarding classification algorithms, the most widely used in the literature consist of support vector machines (SVM) [16, 17, 21–29, 31, 34, 36, 38–41, 43], since these present a good generalization capability, performing well in high dimensional spaces as in neuroimaging data classification. Other strategies have been considered, including k-nearest neighbors (k-NN) [16], linear discriminant analysis (LDA) and Gaussian discriminant analysis (GDA) [20, 30]. Several studies have used ensemble classifiers (including random forests and extremely randomized trees [18, 42]), where multiple "weak" classifiers are combined to improve the generalization ability and robustness of the model, reducing possible overfitting problems [38]. Regarding classification between AD and cognitively normal (CN) subjects using image registration, accuracies of 95.7% have been attained [39], while reported values for sensitivity (specificity) are 98.78% [30] (98.2% [39]). As for the classification between CN and MCI, the corresponding highest reported values are, respectively, 92.36% [19], 99.58% [38] and 90.40% [21]. For studies that did not apply the image registration step, or only did so in the training stage, the highest accuracy presented for CN vs. AD was, respectively, 89.7% [33] (with 88.0% sensitivity and 92.6% specificity) and 92.75% [32] (with 93.48% sensitivity and 93.50% specificity). In [33], the performance of the CAD system in the CN vs. MCI problem was also evaluated, reaching accuracy, sensitivity and specificity values of, respectively, 74.0%, 86.6% and 92.6%.

3. Theoretical Framework

3.1. Feature transformation

The simplest choice of features for image classification consists of raw voxel intensities, a direct measure of fluorodeoxyglucose (FDG) uptake (for the purpose of this work) detected in each voxel. Additional alternatives were also considered, namely histogram of textons and stacked sparse autoencoder (SSAE) feature representations.

Histogram of textons

In this bag-of-features approach, the texture of the images to be classified is considered. As these have spatially repeating properties, there should be several distinct filter response vectors together with noisy variations of them, allowing for clustering these into a small set of prototypes, designated as textons, using an algorithm such as k-means clustering [48, 49]. In this work, the sequential version of this algorithm was used, performing data assignment and centroid update one training sample at a time.

After building the texton dictionary from a set of training images, model extraction is performed (depicted in Figure 2 for the 2D case), so that each image is represented by a (normalized) histogram of texton frequencies. The same procedure is applied on the test set [50].

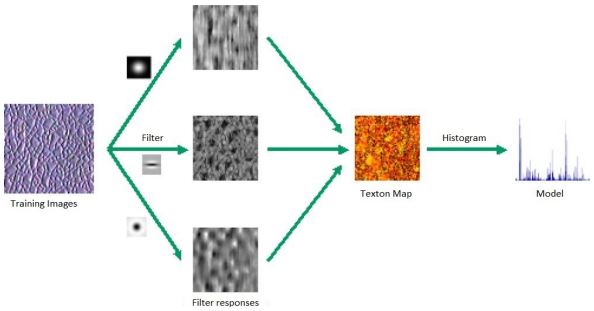

Figure 2: Illustration of the model extraction procedure.

Amongst the filter banks commonly used for this approach are the Maximum Response (MR) sets, where the filters are rotationally invariant, which constitutes an interesting property for this work, as the images used in one particular dataset were allowed to appear at any orientation [50]. In this thesis, an extension to the 3D-MR8 filter bank proposed in [16] was applied, which includes a 3D Gaussian filter and its Laplacian, 3D edge filters, bar and plane filters (at 3 triplets of scales and 61 orientations), as shown in Figure 3 [16].

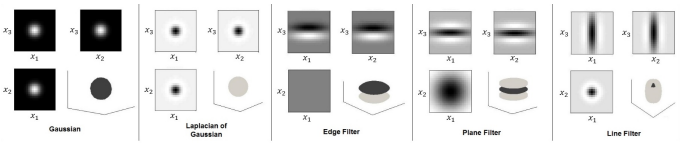

Figure 3: Representation of an example of each type of filter in the 3D extension of the MR8 set.

Some pre-processing steps should be applied for this approach, namely image intensity standardization, filter normalization to unit l_1 norm, and normalization of the filter responses $F(x)$ according to [50].

SSAE feature representations

The SSAE comprises multiple layers of sparse autoencoders in which the outputs of each layer are wired to the inputs of the successive one. The autoncoder, depicted in Figure 4, aims to minimize the discrepancy between input x and reconstruction $\hat{x}$, by learning an encoder and a decoder, yielding a set of weights W and biases b [51,52]. It consists of an unsupervised learning algorithm, as the label information isn't used at any point [53]. Limiting the number of hidden units or enforcing a sparsity constraint, this algorithm is forced to learn a compressed representation of the input, obtaining a low-dimensional, high-level, feature representation [52,53], similar to applying pre-defined filters in the previous texton-based approach.

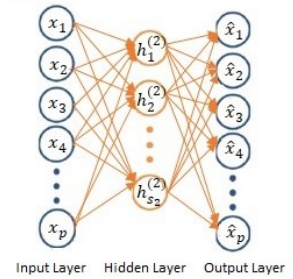

Figure 4: Representation of the architecture of an AE.

Sparse autoencoders result from enforcing the sparsity constraint $\hat{\rho}_j = \rho$, where $\hat{\rho}_j$ is the average activation of hidden unit j and ρ the sparsity parameter, typically close to 0, so that most hidden units' activations must be near 0 [52]. The overall cost function considers, thus, the deviation between input and reconstruction, a regularization term to help prevent overfitting (using weight decay [51]), and one to penalize $\hat{\rho}_j$ deviating significantly from ρ, based on the Kullback-Leibler (KL) divergence [52].

As for the stacked SAE, a greedy layer-wise training approach can be used, where the network layers are trained one at a time, sequentially performing forward passes and backpropagation [37]. After obtaining the optimal parameters, the SSAE transforms the input feature vector to a new feature representation [53]. Particularly for this work, it transforms the voxel intensities of each input patch to its new feature representation to be further classified.

3.2. Classification

In the supervised learning methods proposed, SVM, naive Bayes and the softmax classifier were employed. SVM was used due to its good generalization capability when handling high dimensional data, while naive Bayes was chosen since its computational cost is considerably lower and, from its intrinsic assumption, the order in which the features are fed into the classifier are irrelevant, an interesting property for this work as the images are not registered and so the extracted features may be encountered in different orders. Regarding SSAE, a softmax classifier was applied.

Support vector machines

The SVM algorithm for classification aims to maximize the distance between the decision boundary and the closest training patterns (the support vectors), where the solution hyperplane is derived from a quadratic programming optimization problem [54, 55]. In the simplest, linearly separable case, the following constraint is satisfied [55]:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \ , \quad y_i \in \{-1, 1\}, \quad \forall i \tag{1}$$

The algorithm aims to maximize $d = \frac{2}{\|w\|}$, so the optimization problem can be formulated as [54, 55]:

$$\min \frac{1}{2} \|w\|^2 \ , \quad s.t. \ \ y_i(x_i \cdot w + b) - 1 \geq 0 \ , \quad \forall i \tag{2}$$

and the decision function is given by:

$$\hat{y} = f(x) = sign(x \cdot w + b), \quad \hat{y} \in \{-1, 1\} \tag{3}$$

To deal with non-separable data, data points are allowed on the wrong side of the hyperplane provided that they suffer a penalty, so that positive slack variables ξ_i are introduced and the optimization problem becomes [54–56]:

$$\begin{aligned} &\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \ , \\ &s.t. \ \ y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \ , \quad \forall i \end{aligned} \tag{4}$$

hence switching to a soft margin. There is a trade-off between these and hard margins, as the former allows for errors in the training set but helps prevent overfitting [56].

Since only the inner product between input patterns is required for computing the separating hyperplane, when the decision boundary can't be synthesized as a linear function of the data, the problem can be solved applying the kernel trick [54–57]. Available kernel functions include linear, radial basis function (RBF) and generalized histogram intersection (GHI), as given by:

$$K_{Linear}(x_i, x_j) = x_i^T x_j \tag{5}$$

$$K_{RBF}(x_i, x_j) = e^{-\frac{1}{2\sigma^2}\|x_i - x_j\|^2} \tag{6}$$

$$K_{GHI}(x_i, x_j) = \sum_{k=1}^{p} min(|x_{ik}|^\beta, |x_{jk}|^\beta) \tag{7}$$

where x_{ik} corresponds to the k-th component of the i-th training sample and σ and β are hyperparameters of the model [16, 54, 55, 58].

Naive Bayes

In general terms, the Bayes classifier builds upon the Bayes law and assigns an input pattern with feature vector x the class for which the *a posteriori* probability $P(\omega_i|x)$ is the highest, as given by [59–61]:

$$\hat{y} = arg \max_{\omega \in \Omega} P(\omega|x) \tag{8}$$

The naive Bayes derives from making a strong independence assumption, particularly that all the features are conditionally independent given the class, yielding [59–61]:

$$P(\omega_k|x) = \frac{\prod_{i=1}^{p} P(x_i|\omega_k)}{P(x)} \times P(\omega_k) \tag{9}$$

where the normalization term $P(x)$ isn't required for classification as it isn't class dependent. Even though the conditional independence assumption imposed in naive Bayes is unrealistic, turning into a suboptimal classifier if it's false, it often leads to good results [60].

Softmax classifier

Concerning the softmax classifier, it generalizes the logistic regression function to multi-class problems and it's typically used in the final layer of deep neural network architectures, including SSAE. Since it corresponds to a normalized exponential function, the output value can be used as an estimator of the conditional probability that the input pattern belongs to each class, attributing it to the one that maximizes this value [37].

4. Implementation

4.1. Datasets

Three FDG-PET datasets were used for each class, as depicted in Figure 5, two of which obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, and the third generated from those. The (real) non-registered dataset comprises co-registered, averaged, standardized images and voxel sizes (reoriented into a 160x160x96 grid) with uniform resolution, not having undergone non-linear warping nor linear scaling, contrary to the registered images, non-linearly warped to the Talairach brain atlas of 128x128x60 grid [62]. To obtain the artificially genereated dataset, 3° to 309° rotations along the inferior/superior axis were applied on each respective registered image. By imposing a CDR score of 0, 0.5 and 0.5 or higher, respectively for the CN, MCI and AD class, 50, 80 and 53 patients were selected for each.

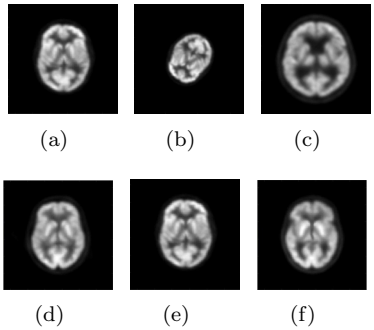

Figure 5: The (a) registered, (b) generated and (c) non-registered datasets, for (d) CN, (e) MCI and (f) AD.

4.2. Experimental setup

In the texton-based approach, an extension to the 3D-MR8 introduced in 3.1 was used, composed by Gaussian, Laplacian of Gaussian (LoG), edge, bar and plane filters, included at 3 triplets of scales: $(\sigma_x, \sigma_y, \sigma_z) = \{(1,1,1); (2,2,2); (4,4,4)\}$ for Gaussian and LoG filters; $(\sigma_x, \sigma_y, \sigma_z) = \{(1.5, 1.5, 0.5); (3,3,1); (6,6,2)\}$ for the edge and plane filters; and $(\sigma_x, \sigma_y, \sigma_z) = \{(0.5, 0.5, 1.5); (1,1,3); (2,2,6)\}$ for the bar filter. The last 3 were replicated with multiple orientations through systematic sampling of the angles around z and y, ranging from $\frac{\pi}{6}$ to $\frac{5\pi}{6}$, and 0 to $\frac{11\pi}{6}$, respectively, from which 9 filter responses were used as to achieve rotation invariance. Combined with those from Gaussian and LoG filters, a 15-dimensional filter response vector was obtained. The pre-processing steps presented in 3.1 were also applied.

A grid-search procedure was used to tune the hyperparameters associated with the linear, RBF and GHI kernels through nested cross-validation, using the *LIBSVM* toolbox [63] compatible with *MATLAB R2016a*. 10 folds were used, avoiding class imbalance, and the allowed hyperparameter values were such that $C \in \left\{2^{-10}, 2^{-8}, ..., 2^{10}\right\}$, $\beta \in \{0.1, 0.4, ..., 1.9\}$ and $\gamma \in \left\{2^{-11}, 2^{-9}, ..., 2^{3}\right\}$, where γ corresponds to $\frac{1}{2\sigma^2}$ in Equation 6. As for naive Bayes, to account for the possible occurrence of negligible texton frequencies in the extracted histograms, the log-probability of occurrence of each feature in either class was considered and Laplace (additive) smoothing was introduced.

As for SSAE, the *Neural Network Toolbox* of *MATLAB R2016a* was used. The first SAE (with 100 hidden units, trained for 400 epochs maximum) was subjected to two regularization techniques, namely an L2 regularizer for the network weights (with 0.004 coefficient) and a sparsity regularizer (with impact parameter 4 and ρ of 0.15, defined in 3.1). Analogous parameters were set for the second SAE (with 50 hidden units, trained for 100 epochs maximum), respectively 0.002, 4 and 0.1. The final softmax classifier was trained for 400 epochs maximum and the whole stacked neural network was fine tuned using backpropagation with scaled conjugate gradient (SCG) [64], for 100 epochs maximum. The performance function for the SAEs was the mean squared error (MSE), while cross-entropy was used for supervised learning.

4.3. Proposed approaches

For each binary classification task, two feature extraction approaches were considered, namely using the whole brain, and patches of the brain. This was chosen since, when disregarding image registration, neither the most discriminative voxels nor ROIs could be directly selected, as its location would vary, while no prior identification was required if the features were extracted from the whole brain or patches appropriately selected for each subject.

To exclude the area surrounding the brain, a mask was used. Regarding the registered dataset, this consisted of a pre-defined $128 \times 128 \times 60$ mask in the Talairach space, as depicted in Figure 6, while in the non-registered dataset it was created from averaging the brain scans of all subjects, and keeping the voxels for which the resulting intensity value was above 0.5 of the maximum, followed by a morphological operation to fill in small regions that were visually identified as a part of the whole brain area. As for the generated dataset, the mask for each subject was obtained from the same affine transform as for the respective brain image, applied on the Talairach brain mask considered.

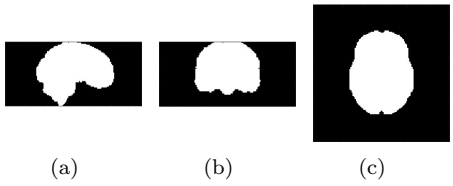

Figure 6: Depiction of (a) sagittal, (b) coronal and (c) axial sections of the brain mask applied in the registered dataset.

Feature extraction from the whole brain

In this case, both SVMs (with linear, GHI and RBF kernels) and naive Bayes were applied, using the voxel intensitiy feature vectors (with SVM) or the texton-based approach. Only the second approach was tested on the generated dataset, as the images produced and respective brain masks had different dimensions, so that the extracted features could not be fed into the classifier without further transformation, as made possible using the texton-based approach. Several experiments were performed regarding the number of textons to use for building the dictionary, ranging from 15 to 10000, employing sequential k-means (as worse performances were achieved by the original k-means clustering algorithm).

Feature extraction from patches of the brain

In contrast to the previous method, here, the feature vectors to be fed into the classifiers could not be raw voxel intensities as these would depend on the order on which the features were fed and consequently on the anatomical position from which the patches were drawn, not being robust to skipping image registration. The features considered for this section were thus either histograms of textons or feature representations learned from an SSAE.

The size of the 3D patches was fixed at 7^3, and the overlap between them was set at 75%, considering the trade-off between computational cost and coverage of the different regions of the brain, as higher overlap would result in increased number of patches and computational cost.

Patches within ROIs

Unlike the previous approach, this strategy was considered in the case that image registration would still be performed in training, which, not being ideal, still allows for avoiding this step when applying the CAD system to diagnose novel subjects, saving computational costs. In the training stage, a patch was labeled as discriminative if it spanned over one or multiple ROIs [16] as shown in Figure 7, including lateral temporal, mesial temporal, inferior frontal gyrus and orbitofrontal, inferior anterior cingulate, dorsolateral parietal, superior anterior cingulate, posterior cingulate and precuneous.

(a) (b) (c)

Figure 7: Depiction of (a) sagittal, (b) coronal and (c) axial sections of the labeled ROIs in the Talairach brain atlas [16]

Also, as the image resolution differed between registered and non-registered datasets, to evaluate the performance of the classifier on the real non-registered dataset, the ROIs had to be manually identified from the ones in the Talairach brain atlas, so that this dataset could also be used in training. A patch was labeled as discriminative if over 90% of its volume was within the reunion of these ROIs, and a SSAE combined with softmax classifier was trained to learn this in the test set. The selected patches were then used for diagnosis, either with a SSAE and output softmax layer or computing the histogram of textons of the selected patches' reunion, followed by applying naive Bayes or SVM (with linear, RBF and GHI kernels), considering a dictionary of 1250 textons.

Patches containing discriminative textons

In this approach, the most discriminative textons were defined as those with higher mutual information ranking. Since the number of textons to consider was unknown, it was optimized through nested cross-validation, ranging between 1, 5, 10 or 15. A patch was selected if its distribution contained, thus, any of these most discriminative textons and final diagnosis resulted from applying SVM (with linear, GHI and RBF kernels) or naive Bayes.

Random patch selection

Random patch selection was also evaluated, selecting as many as the discriminative ones identified in the ROI-based strategy, particularly 1004, 1250 and 5836 for, respectively, the registered, generated and non-registered datasets. Final classification was again attained using SVM (with linear, GHI and RBF kernels) and naive Bayes.

5. CN vs. AD - Results and Discussion

5.1. Whole brain

The final model with the selected hyperparameters was applied on each respective test set, and the best diagnostic accuracy results attained are summarized in Tables 1-3.

Table 1: Registered dataset's best results for CN vs. AD, using whole brain features. Format: Mean (SEM) [%].

Registered Dataset	Accuracy [%]	N°textons
VI - Linear	84.36 (4.02)	-
VI - GHI	86.18 (3.43)	-
Textons - Linear	79.45 (3.40)	300
Textons - GHI	85.36 (3.92)	5000
Textons - RBF	87.27 (3.32)	5000
Textons - N. Bayes	79.64 (3.63)	7500

Table 2: Generated dataset's best results for CN vs. AD, using whole brain features. Format: Mean (SEM) [%].

Generated Dataset	Accuracy [%]	N°textons
Textons - Linear	78.55 (3.32)	200
Textons - GHI	85.36 (3.31)	7500
Textons - RBF	86.09 (2.95)	90
Textons - N. Bayes	75.45 (5.35)	7500

Table 3: Non-registered dataset's best results for CN vs. AD, using whole brain features. Format: Mean (SEM) [%].

Non-Reg. Dataset	Accuracy [%]	N°textons
VI - Linear	76.64 (2.68)	-
VI - GHI	76.73 (4.12)	-
Textons - Linear	69.91 (3.52)	1000
Textons - GHI	76.82 (4.50)	1500
Textons - RBF	80.64 (4.92)	2500
Textons - N. Bayes	60.09 (4.75)	7500

The texton-based approach proved to be more robust to skipping the image registration step, as the raw voxel intensity should be used when each particular feature refers to the same anatomical position across all subjects, which is not true in non-registered datasets. The former can, thus, prove to be more adequate for this application as the model extracted for each image is a histogram of textons, where information of the anatomical structure from where each individual feature originated is disregarded. As for the fact that in the generated dataset the highest diagnostic accuracy reached was very similar to that of the registered dataset and better than that of the non-registered one, it might be due to the dimension of the images being much higher for the (real) non-registered dataset, so that an increased number of textons could be required to enhance the classifier's performance, as the task of separating the two classes is more difficult.

Also, while naive Bayes was slightly outperformed by SVM in all datasets, since its complexity and computational cost is significantly lower to that of the SVM, it remains a good alternative, as it can still perform well and is again robust to skipping the image registration step.

5.2. Patches

Random patch selection

Regarding the use of random patches of the brain, the results are presented in Table 4.

Table 4: Diagnostic accuracy for CN vs. AD, using randomly selected patches. Format: Mean (SEM) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	79.45 (5.56)	81.27 (4.84)	70.00 (4.92)
SVM - GHI	79.55 (3.76)	80.55 (5.27)	69.00 (4.95)
SVM - RBF	77.64 (4.64)	78.55 (4.16)	69.18 (5.75)
N. Bayes	68.64 (5.71)	73.55 (5.36)	66.82 (4.96)

It should be noted that, even though the patches were selected completely randomly, the classifiers could still perform well, which might be due to the fact that a sufficient number of patches was selected, so that many of these could be drawn from regions that are relevant for the diagnosis. Similar performances were reported for SVM and naive Bayes, again highlighting the advantages of the latter, for which the computational cost is extremely smaller. The fact that this procedure is completely random can also explain why better classification results could be reached in the generated dataset than in the registered one, so that it can be considered to be robust to skipping image registration.

Patches containing discriminative textons

Concerning the selection of patches containing discriminative textons, the results are presented in Table 5.

Table 5: Diagnostic accuracy for CN vs. AD, using patches with discriminative textons. Format: Mean (SEM) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	77.45 (5.29)	67.00 (5.15)	66.82 (5.79)
SVM - GHI	75.36 (5.22)	70.91 (4.91)	65.09 (5.25)
SVM - RBF	72.55 (5.11)	71.91 (3.36)	68.18 (5.29)

While these results are reasonable, these models don't outperform the previous seed for random patch selection, so that further methods should be considered and explored. In general, however, in theory this should attain similar performances regardless of the image registration step, as it doesn't require any knowledge on the anatomical position from which the patches are drawn and there is significant overlap between these.

Patches within ROIs

Concerning patch selection, the results from employing SSAE to learn feature representations of the input patches, followed by the output softmax classifier layer, are presented in Table 6. It can be observed that the accuracy in patch selection using this method was very high regarding the registered dataset, while worse performances were achieved for the non-registered one, as would be expected being the latter a more difficult classification problem. Moreover, since the final binary classification between CN and AD must be preceded by this step, the error obtained regarding patch selection propagates to that that final classification, as the diagnosis is then performed using patches within regions that might not be affected by the disease, possibly reducing the sensitivity of the method.

Table 6: Patch selection accuracy for CN vs. AD, considering previously identified ROIs. Format: Mean (SEM) [%].

	Registered [%]	Non-registered [%]
SSAE+Softmax	91.18 (0.13)	79.74 (0.23)

The selected patches were then fed as input to the final classification algorithm between CN and AD. Several methods were explored in this step, namely using again an SSAE and softmax classifier at the output layer where each patch was first classified in either CN or AD, followed by majority voting to obtain the final diagnosis. Alternatively, the histogram of textons on the reunion of all selected patches was also computed and fed into an SVM or naive Bayes classification algorithm. Another method consisted of again attaining the final diagnosis through majority voting in combination with the histogram of textons computed for each patch. Nonetheless, due to the extremely high computational cost associated to the nested cross-validation procedure for optimization of the hyperparameters of the SVM algorithm, and due to the fact that the final diagnostic accuracy could largely depend on these parameters such that fixing these could also be misleading, and that naive Bayes could also not perform well in this task, as initially tested, these results were excluded from this discussion and hence not presented in Table 7.

Table 7: Diagnostic accuracy for CN vs. AD, considering the selected patches within ROIs. Format: Mean (SEM) [%].

	Registered [%]	Non-registered* [%]
SVM - Linear (Reunion)	81.45 (3.50)	76.36 (3.64)
SVM - GHI (Reunion)	76.36 (4.84)	76.36 (3.64)
SVM - RBF (Reunion)	77.36 (4.27)	76.36 (3.64)
N. Bayes (Reunion)	62.62 (4.68)	71.82 (3.66)
SSAE+Softmax (Maj. Voting)	89.09 (4.07)	85.45 (5.45)

It can be observed that SSAE followed by softmax performed very well on the registered dataset, reaching a diagnostic accuracy close to the state of the art (around 90%), while the texton-based approach could also lead to good results, namely using the linear kernel. It is important to highlight that, due to the high computational cost associated to training the full network, the analysis regarding the non-registered dataset was performed using only 2 (and not 10) folds (as indicated by * in Table 7), so that the results here displayed for the latter dataset, despite being quite satisfactory, could be positively biased.

6. CN vs. MCI - Results and Discussion

6.1. Whole brain

The selected model was applied on the respective test set (for each dataset), obtaining the final diagnostic accuracy results for this binary task summarized in Tables 8-10.

Similar to the previous task, the performance of most models would improve when considering larger numbers of textons in the dictionary, so that it could be hypothesized that better diagnostic accuracy results could have been attained if this was extended beyond 10000; this should also

be particularly the case for the non-registered dataset, as the dimensions of these images are much greater. However, this would require a higher computational cost, resulting in a trade-off between the two.

Table 8: Registered dataset's best results for CN vs. MCI, using whole brain features. Format: Mean (SEM) [%].

Registered Dataset	Accuracy [%]	N°textons
VI - Linear	62.31 (4.50)	-
VI - GHI	63.85 (4.87)	-
Textons - Linear	67.69 (2.06)	500
Textons - GHI	73.08 (3.85)	10000
Textons - RBF	70.77 (3.20)	10000
Textons - N. Bayes	67.69 (2.99)	10000

Table 9: Generated dataset's best results for CN vs. AD, using whole brain features. Format: Mean (SEM) [%].

Generated Dataset	Accuracy [%]	N°textons
Textons - GHI	63.08 (3.20)	7500
Textons - RBF	64.62 (3.84)	7500
Textons - N. Bayes	66.92 (4.14)	7500

Table 10: Non-registered dataset's best results for CN vs. AD, using whole brain features. Format: Mean (SEM) [%].

Non-Reg. Dataset	Accuracy [%]	N°textons
VI - Linear	61.54 (3.24)	-
Textons - GHI	61.54 (3.63)	7500
Textons - N. Bayes	60 (3.40)	10000

6.2. Patches

Random patch selection

Regarding the use of random patches of the brain, the results are presented in Table 11.

Table 11: Diagnostic accuracy for CN vs. MCI, using randomly selected patches. Format: Mean (SEM) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	54.62 (4.36)	43.08 (3.48)	39.23 (0.77)
SVM - GHI	66.92 (3.81)	55.38 (2.99)	56.92 (3.84)
SVM - RBF	64.62 (4.17)	53.08 (5.06)	42.31 (3.85)
N. Bayes	62.31 (3.13)	62.31 (5.19)	59.23 (3.45)

As in the previous classification task, naive Bayes performed similarly to SVM, while significantly reducing the complexity and computational cost of the problem. However, the poor performances observed are close to a random procedure for binary classification.

Patches containing discriminative textons

Concerning the selection of patches containing discriminative textons, the results are presented in Table 12.

Table 12: Diagnostic accuracy for CN vs. MCI, using patches with discriminative textons. Format: Mean (SEM) [%].

	Registered [%]	Generated [%]	Non-registered [%]
SVM - Linear	56.15 (3.45)	49.23 (3.66)	47.69 (3.40)
SVM - GHI	66.92 (3.04)	60.77 (4.51)	62.31 (3.71)
SVM - RBF	57.69 (3.29)	49.23 (4.17)	44.62 (8.01)

The applied algorithms for this approach outperform the results obtained with the previous seed for random patch selection slightly, but further techniques should be considered and explored, although, as mentioned, this should be able to exhibit similar performances regardless of the image registration step.

Patches within ROIs

As in the previous binary classification problem, for this approach two steps were performed, namely patch selection and final image classification. The attained results for patch selection are presented in Table 13, where it can be observed that while the SSAE with softmax classifier performed very well in the registered dataset, worse performances were achieved for the non-registered one, being theoretically a more complicated classification problem.

Table 13: Patch selection accuracy [%] for CN vs. MCI, considering previously identified ROIs. Format: Mean (SEM).

	Registered [%]	Non-registered [%]
SSAE+Softmax	91.55 (0.12)	79.92 (0.12)

As in the CN vs. AD classification task, the selected patches were then fed as input to the final classification algorithm between CN and MCI. Several methods were explored in this step, as presented in Table 14.

Table 14: Diagnostic accuracy for CN vs. MCI, considering the selected patches within ROIs. Format: Mean (SEM) [%].

	Registered [%]	Non-registered* [%]
SVM - Linear (Reunion)	73.85 (4.47)	42.31 (3.85)
SVM - GHI (Reunion)	74.62 (3.98)	61.54 (3.44)
SVM - RBF (Reunion)	77.69 (4.51)	42.31 (3.85)
N. Bayes (Reunion)	65.38 (2.63)	46.15 (0)
SSAE+Softmax (Maj. Voting)	61.54 (0)	61.54 (0)

It can be observed that the method which performed better was the linear SVM applied to the reunion of selected patches, for the registered dataset, while the SSAE followed by softmax and majority voting led to the highest classification accuracy in the non-registered dataset. Once again, it should be higlighted that only 2 folds (* in Table 7) were used in the non-registered dataset, so that its results could eventually be positively biased. Also, the propagation of error in patch selection might have led to worse performances, particularly in the non-registered dataset.

6.3. Summary

Regarding CN vs. AD, the best accuracy results obtained were 89.09%, 86.09% and 80.63%, respectively for the registered, generated and non-registered datasets, the former using patch selection and majority voting by means of SSAE and the softmax classifier, and the remainder using textons extracted from the whole brain images, which proved to be robust to registration errors. High sensitivity (specificity) values were also obtained, with a maximum of 90% (88%), 86% (86%) and 77.33% (88%), respectively for the registered, generated and non-registered datasets.

Concerning the CN vs. MCI problem, being a much more complicated classification task, the attained results in terms of both accuracy, sensitivity and specificity were naturally lower. Regarding accuracy, values of 77.69%, 66.92% and 62.31% were respectively reported for the registered, generated and non-registered datasets, the former

using patch selection and further classification using histogram of textons on the reunion of all patches, while the remainder using textons extracted from the whole brain images and patches containing discriminative textons, respectively. For the former method, a value of 100% was reported for the sensitivity in the registered dataset, although at the cost of a 50% specificity, and so the method's performance was not satisfactory. Particularly this corresponds to the case where all MCI patients were correctly classified, while all the cognitively normal subjects were misclassified as having the disease (false alarms), meaning that in fact all subjects were classified as having MCI. Reasonable specificity values, close to 60%, were also reported for the majority of the methods explored in this classification problem. Nonetheless, considering that it might be more relevant to accurately predict the diagnosis for subjects having MCI or AD, the results obtained for both methods might in this sense be considered satisfactory, as the majority of these could perform better considering sensitivity than specificity.

7. Conclusions

In what refers to the achievements of this work, as the classification algorithms that used both the texton-based approach with features extracted from the whole brain and the learned feature representations using the stacked sparse autoencoder could perform well in each binary classification problem (although more evidently so for the CN vs. AD task), regardless of the application or not of the image registration step, it can be considered that its major objective, of obtaining a method for CAD of AD robust to registration errors, was achieved.

Concerning possible approaches to be tested in the future, as mentioned throughout this work, it would be of interest to study how switching from using the 3D-MR8 filter bank to a set of filters (weights) to be learned by a stacked sparse autoencoder would influence the performance of the models in each classification problem, when using the same texton-based approach. Further studies could also be performed to expand the results presented in this work, namely regarding the influence of the size of the constructed patches, or else performing majority voting on the patches selected through SSAE combined with the softmax classifier, or also considering its application to the whole brain images. Other deep learning strategies that have attained state of the art performances could be explored too, namely combining autoencoders with convolutional neural networks, or else considering deep Boltzmann machines, amongst many other promising techniques.

Acknowledgements

I would like to thank Professor Margarida Silveira for all the useful remarks and guidance, and my friends and family for the constant support.

References

[1] Alzheimer's Association et al. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018.

[2] T.K. Khan. Chapter 2 - Clinical diagnosis of Alzheimers disease. In T.K. Khan, editor, *Biomarkers in Alzheimer's Disease*, pages 27 – 48. Academic Press, 2016.

[3] Alistair Burns and Steve Iliffe. Alzheimer's disease. *BMJ*, 338, 2009.

[4] Denham Harman. Alzheimer's disease pathogenesis. *Annals of the New York Academy of Sciences*, 1067(1):454–460, 2006.

[5] Giovanni B Frisoni et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *The Lancet Neurology*, 16(8):661–676, 2017.

[6] Dennis J Selkoe. The molecular pathology of Alzheimer's disease. *Neuron*, 6(4):487–498, 1991.

[7] Reisuke H Takahashi et al. Plaque formation and the intraneuronal accumulation of β-amyloid in Alzheimer's disease. *Pathology international*, 67(4):185–193, 2017.

[8] Judith Neugroschl and Sophia Wang. Alzheimer's disease: diagnosis and treatment across the spectrum of disease severity. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 78(4):596–612, 2011.

[9] Claudia Woolf et al. Can the clinical dementia rating scale identify mild cognitive impairment and predict cognitive and functional decline? *Dementia and geriatric cognitive disorders*, 41(5-6):292–302, 2016.

[10] Y-L Chang et al. Global clinical dementia rating of 0.5 in MCI masks variability related to level of function. *Neurology*, 76(7):652–659, 2011.

[11] John C Morris. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 1993.

[12] Keith A Johnson et al. Brain imaging in Alzheimer's disease. *Cold Spring Harbor perspectives in medicine*, page a006213, 2012.

[13] Karl Herholz et al. Positron emission tomography imaging in dementia. *The British journal of radiology*, 80(special_issue_2):S160–S167, 2007.

[14] Takashi Kato et al. Brain fluorodeoxyglucose (FDG) pet in dementia. *Ageing research reviews*, 30:73–84, 2016.

[15] Yangming Ou et al. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. *IEEE transactions on medical imaging*, 33(10):2039–2065, 2014.

[16] Pedro Morgado et al. Texton-based diagnosis of Alzheimer's disease. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6. IEEE, 2013.

[17] Imene Garali et al. Region-based brain selection and classification on PET images for Alzheimer's disease computer aided diagnosis. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1473–1477. IEEE, 2015.

[18] Marie Wehenkel et al. Computer aided diagnosis system based on random forests for the prognosis of Alzheimers disease. In *1st HBP Student Conference-Transdisciplinary Research Linking Neuroscience, Brain Medicine and Computer Science*. Frontiers Media SA, 2018.

[19] Adrien Payan and Giovanni Montana. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*, 2015.

[20] Pierrick Coupé et al. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: clinical*, 1(1):141–152, 2012.

[21] Manhua Liu et al. Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.

[22] Manhua Liu et al. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Human brain mapping*, 35(4):1305–1319, 2014.

[23] Laila Khedher et al. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, 151:139–150, 2015.

[24] Yudong Zhang et al. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*, 9:66, 2015.

[25] Mohamed M Dessouky et al. Computer-aided diagnosis system for Alzheimers disease using different discrete transform techniques. *American Journal of Alzheimer's Disease & Other Dementias®*, 31(3):282–293, 2016.

[26] C Jongkreangkrai et al. Computer-aided classification of Alzheimer's disease based on support vector machine with combination of cerebral image features in MRI. In *Journal of Physics: Conference Series*, volume 694, page 012036. IOP Publishing, 2016.

[27] Jun Zhang et al. Detecting anatomical landmarks for fast Alzheimers disease diagnosis. *IEEE Transactions on Medical Imaging*, 35(12):2524–2533, 2016.

[28] Iman Beheshti et al. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in biology and medicine*, 83:109–119, 2017.

[29] Laila Khedher et al. Independent component analysis-support vector machine-based computer-aided diagnosis system for Alzheimers with visual support. *International journal of neural systems*, 27(03):1650050, 2017.

[30] Chen Fang et al. A novel Gaussian discriminant analysis-based computer aided diagnosis system for screening different stages of Alzheimer's disease. In *Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on*, pages 279–284. IEEE, 2017.

[31] Ramesh Kumar Lama et al. Diagnosis of Alzheimers disease based on structural MRI images using a regularized extreme learning machine and PCA features. *Journal of healthcare engineering*, 2017, 2017.

[32] Mingxia Liu et al. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

[33] Fan Li et al. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 2018.

[34] Javier Ramírez et al. Computer-aided diagnosis of Alzheimers type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72, 2013.

[35] Olfa Ben Ahmed et al. Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. *Neurocomputing*, 220:98–110, 2017.

[36] Esther E Bron et al. Multiparametric computer-aided differential diagnosis of Alzheimers disease and frontotemporal dementia using structural and advanced MRI. *European radiology*, 27(8):3372–3382, 2017.

[37] Siqi Liu et al. Early diagnosis of Alzheimer's disease with deep learning. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 1015–1018. IEEE, 2014.

[38] Heung-Il Suk et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[39] Xiaofeng Zhu et al. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Medical image analysis*, 38:205–214, 2017.

[40] Tien Duong Vu et al. Multimodal learning using convolution neural network and sparse autoencoder. In *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on*, pages 309–312. IEEE, 2017.

[41] Arthur Mikhno et al. Multimodal classification of dementia using functional data, anatomical features and 3D invariant shape descriptors. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 606–609. IEEE, 2012.

[42] Tong Tong et al. Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern recognition*, 63:171–181, 2017.

[43] Yue Cui et al. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. 6(7):e21896, 2011.

[44] Pedro Morgado et al. Diagnosis of Alzheimer's disease using 3D local binary patterns. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 1(1):2–12, 2013.

[45] Wilkin Chau and Anthony R McIntosh. The Talairach coordinate of a point in the MNI space: how to interpret it. *Neuroimage*, 25(2):408–416, 2005.

[46] DL Collins et al. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192–205, 1994.

[47] MRC Cognition and Brain Sciences Unit. University of Cambridge. The MNI brain and the Talairach atlas, 2009. Last access: September 28th, 2018.

[48] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[49] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44, 2001.

[50] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *International journal of computer vision*, 62(1-2):61–81, 2005.

[51] Jun Xu et al. Stacked sparse autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 999–1002. IEEE, 2014.

[52] Yao Ju et al. A deep learning method combined sparse autoencoder with SVM. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on*, pages 257–260. IEEE, 2015.

[53] Jun Xu et al. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016.

[54] Armin Shmilovici. Support vector machines. In *Data mining and knowledge discovery handbook*, pages 231–247. Springer, 2009.

[55] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[56] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[57] Bernhard E Boser et al. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[58] Sabri Boughorbel et al. Generalized histogram intersection kernel for image recognition. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–161. IEEE, 2005.

[59] David D Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[60] Nir Friedman et al. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[61] Irina Rish et al. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[62] ADNI. Alzheimer's Disease Neuroimaging Initiative, 2017. Accessed: 25/04/2018.

[63] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[64] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.